介紹

本次實作為研究學生的成績與父母教育程度、家中經濟及有無預習是否有正相關性，及假設有之預測結果。

資料集介紹

本次使用之資料及為 KAGGLE 上的 STUDENT PERFORMANCE IN EXAM，其中分別有性別、父母教育程度、午餐、數學成績、組別、有無做準備課程、閱讀成績、寫作成績。

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	male	group D	some high school	free/reduced	completed	71	69	67
1	female	group A	bachelor's degree	standard	none	59	72	76
2	female	group D	high school	standard	none	63	61	64
3	female	group C	bachelor's degree	free/reduced	none	50	55	52
4	female	group D	master's degree	standard	none	85	95	97

資料預處理

將本次用不到的資料移除，並檢查有無空值。

```
df.isnull().sum() #確認無空值
```

```
gender                         0
race/ethnicity                 0
parental level of education    0
lunch                          0
test preparation course        0
math score                     0
reading score                  0
writing score                  0
dtype: int64
```

```
del df['gender']
del df['race/ethnicity']
df.head() #移除不需要的欄位
```

	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	some high school	free/reduced	completed	71	69	67
1	bachelor's degree	standard	none	59	72	76
2	high school	standard	none	63	61	64
3	bachelor's degree	free/reduced	none	50	55	52
4	master's degree	standard	none	85	95	97

分別算出各項分數是否及格(>60)並新增全部及格欄位

```
df['OverAll_PassStatus'] = df.apply(lambda x : 'F' if x['Math_PassStatus'] == 'F' or
                                    x['Reading_PassStatus'] == 'F' or x['Writing_PassStatus'] == 'F' else 'P', axis =1)

df.OverAll_PassStatus.value_counts() #找出全部測驗及格人數
```

```
P    597
F    403
Name: OverAll_PassStatus, dtype: int64
```

將所有需要用到的欄位量化

```
textToInt = {'some high school':1,'high school':2,'some college':3,"associate's degree":4,"bachelor's degree":5,"master's degree"
df['edu level'] = df['parental level of education'].map(textToInt)
changeToInt = {'completed':1,'none':0}
df['prepared']=df['test preparation course'].map(changeToInt)
getPass = {'P':1,'F':0}
df['Pass']=df['OverAll_PassStatus'].map(getPass)
df.head() #程度量化
```

	parental level of education	lunch	test preparation course	math score	reading score	writing score	OverAll_PassStatus	Total_Marks	Percentage	Grade	edu level	prepared	Pass
0	some high school	free/reduced	completed	71	69	67	P	207	69.000000	E	1	1	1
1	bachelor's degree	standard	none	59	72	76	F	207	69.000000	F	5	0	0
2	high school	standard	none	63	61	64	P	188	62.666667	E	2	0	1
3	bachelor's degree	free/reduced	none	50	55	52	F	157	52.333333	F	5	0	0
4	master's degree	standard	none	85	95	97	P	277	92.333333	B	6	0	1

機器學習或深度學習方法

本次使用的方法為 Logistic Function / Sigmoid Function，從老師提供的範例 SNS_ADS 上修改，詳細請查閱程式碼。

研究結果及討論

經過本次研究發現，無論是家中經濟、父母教育程度或有無試前準備，都與學生成績有正相關，其中影響最大的為家中經濟，經濟較好家庭的小孩比較差的小孩平均成績多了約 12 分，而試前準備課程的參加有無則差距 6 分，教育程度上差距最大的為高中-博士，差距為 11 分，可見這些因素都會導致學生成績，而以此基礎來做回歸預測，卻發現預測正確率只有百分之五十三，推測可能原因為用於預測的 FEATURE 數值波動太低，將資料更細分、增加 FEATURE，或使用別的預測方法可能使預測更為準確。

結論

學生成績與家中經濟狀況、父母教育程度、有無試前準備有正相關。

參考文獻

https://www.kaggle.com/spscientist/student-performance-in-exams

https://seaborn.pydata.org/index.html

老師提供的範例